DOCUMENT RESUME

ED 274 476 RC 015 906

AUTHOR Miller, Michael K.; Farmer, Frank L.
TITLE Substantive Nonadditivity in Rural Sociological Research: A Note on Induced Collinearity and Measurement and Testing of Effects.
PUB DATE Aug 86
NOTE 28p.; Paper presented at the Annual Rural Sociological Society Meeting (Salt Lake City, UT, August 26-30, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Analysis; *Hypothesis Testing; Interaction; *Models; *Research Methodology; Research Problems; *Social Science Research; Sociology; *Statistical Analysis; Statistical Significance
IDENTIFIERS *Collinearity; Variables

ABSTRACT

Theories employed to explain regularities in social behavior often contain explicit or implicit reference to the presence of nonlinear and/or nonadditive (i.e., multiplicative) relationships among germane variables. While such nonadditive features are theoretically important, the inclusion of quadratic or multiplicative terms in structural equations to model such features can cause significant methodological problems. This paper estimates a set of equations and formally examines how the inclusion of quadratic terms and multiplicative interaction terms contribute to the level of collinearity of ill-conditioning of the input data matrix and the precision of the parameter estimates. It also examines how effects of explanatory variables in nonadditive models can be measured and tested for statistical significance. The issues are exemplified by estimating a set of models for age-sex adjusted mortality in the United States in 1980. Results indicate that collinearity may not be as big a problem for linear structural social science models as is often believed. Although collinearity is increased by adding quadratic and/or multiplicative terms, the effects of the collinearity tend to be localized and entail only variables with a common base. Findings suggest that the substantive insight gained from including theoretically appropriate nonlinear and nonadditive terms outweigh the methodological problems they create. (Author/NEC)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED274476

# Substantive Nonadditivity in Rural Sociological Research: A Note on Induced Collinearity and Measurement and Testing of Effects[1]

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Michael K. Miller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SEP 1986
RECEIVED
ERIC/CRESS

Michael K. Miller
Center for Health Policy Research
Department of Community Health
and Family Medicine
University of Florida
Gainesville, Florida

By

Frank L. Farmer
Department of Agricultural
Economics and Rural Sociology
Pennsylvania State University
University Park, Pennsylvania

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

RC015906

[1] Paper Presented at the Annual Meeting of the Rural Sociological Society, Salt Lake City, Utah, August 1986. We would like to thank Kenneth P. Wilkinson, Glenn Isreal, and A.E. Luloff for comments on an earlier version of the manuscript.



BEST COPY AVAILABLE

Abstract

Theories employed to explain regularities in social behavior often contain reference (explicit or implicit) to the presence of nonlinear and/or nonadditive (i.e. multiplicative) relationships among germane variables. While such nonadditive features are theoretically important, the inclusion of quadratic or multiplicative terms in structural equations to model such features can cause significant methodological problems. This paper estimates a set of equations and formally examines how the inclusion of quadratic terms and multiplicative interaction terms contribute to the level of collinearity or ill-conditioning of the input data matrix and the precision of the parameter estimates. Subsequently we examine how effects of explanatory variables in nonadditive models can be measured and tested for statistical significance. The results indicate that collinearity may not be as big a problem for linear structural social science models as is often believed. Further, although collinearity is increased by adding quadratic and/or multiplicative terms, the effects of the collinearity tend to be localized and entail only variables with a common base. The findings suggest the substantive insight gained from including theoretically appropriate nonlinear and nonadditive terms outweigh the methodological problems they create.

# Substantive Nonadditivity in Rural Sociological Research: A Note on Induced Collinearity and Measurement and Testing of Effects

## Introduction

In reviewing the methodological changes that have taken place in the discipline over a half century, Stokes and Miller (1985) note that, while progress in some areas has been made, the prototype research model published in Rural Sociology is still one that assumes the world operates in a simple, linear, additive fashion with no feedbacks or lag effects and with germane structures that are assumed constant over time. Such a naive formulation is inconsistent with any of the theoretical perspectives employed to explain how societal systems or subsystems operate (Walker and Cohen, 1985; Blalock, 1965). Many of the theoretical formulations postulate nonlinear effects of an independent variable on a dependent variable, just as many have expectations that the effect of one independent variable is conditioned by the state or values of one or more other independent variables (Southwood, 1978). In either case, linear additive multiple regression models where each variable is evaluated as a unique, independent hypothesis (i.e. $b_j = 0$) is an inadequate analysis strategy. To appropriately test the relevant hypothesis in such situations, the model must contain synthetic terms such as powers or products. Parameter estimation in models that contain such interaction terms are multiplier in the prediction of the dependent variable. This is necessarily true because inclusion of an interaction term mandates the examination of the simultaneous impact of two or more variables at one time.

The objective of the current paper is to demonstrate that the benefit of theoretical appropriateness realized by the inclusion of



BEST COPY AVAILABLE

nonlinear or nonadditive terms in an empirical model is accompanied by increased complexity in the measurement and statistical testing of causal effects and may be accompanied by deleterious methodological side effects. In particular, the inclusion of terms to account for statistical interaction or nonlinear relationships means that causal impacts or "rate-of-change" effects are no longer measured adequately by the regression coefficient on a single independent variable. Instead, the measure of the rate-of-change in the dependent variable for a change in the independent variable (i.e. the partial derivative) is summarized by some combination of regression coefficients. Further, the magnitude of the causal effect in such formulations are not constant. The rate at which the dependent variable changes as a result of a unit change in the independent variable depends on the values assumed by the independent variable (the nonlinear case) or the value of one or more variables that interact with the independent variable of concern (the nonadditive or interaction case). To test such change measure for statistical significance involves testing the partial derivative across a range of chosen values of the relevant variables. For example, it may be appropriate to evaluate the effect measure at the mean of the sample and at values one and two standard deviations either side of the sample mean. Because the value of the standard error of the partial derivative depends on the value of the independent or control variable under investigation, the effect measure may be statistically significant at some levels and not at others. Hence, it is important to test for statistical significance across a range of values that is theoretically and empirically meaningful.

BEST COPY AVAILABLE

In addition to the increased complexity in the measurement and testing of effects, the inclusion of nonadditive or nonlinear terms in regression models also builds in problems of collinearity. In the current manuscript we examine: (1) What impact the inclusion of product and power terms has on the levels of collinearity, and the implications for the statistical efficiency or precision of the parameter estimates; (2) How change effects are appropriately measured, tested for statistical significance, and interpreted when the model includes a quadratic term and/or a product interaction term. The issues are exemplified by estimating a set of models for age-sex adjusted mortality in the United States in 1980.

Model Specification

There are numerous formulations in the literature to explain differential mortality in the United States. One formulation that is consistent with a macrostructural perspective of stratification and inequality is summarized in the behavioral equation

$$\text{EQ 1:}\quad \text{mortality} = \hat{\beta}_0 + \hat{\beta}_1(\%\ \text{Nonwhite}) + \hat{\beta}_2(\text{Income}) + \hat{\beta}_3(\text{Education}) + \hat{\beta}_4(\%\ \text{Urban}) + \hat{\beta}_5(\%\ \text{White Collar}) + \varepsilon$$

The estimated coefficients for the linear model (Table 1) indicate that the nature of the impacts are, by-in-large, consistent with theoretical expectations. Mortality rates increase as the concentration of nonwhite population increases (Mechanic, 1968) and decline with increasing levels of income (Hadley, 1982) and education (Grossman, 1973, Kitagawa & Houser, 1968, Fuch, 1974). The direction of the impact of occupation is as expected (Ahern, 1979), but not statistically


BEST COPY AVAILABLE

significant, and the impact of rurality/urbanity is equivocal. From this point on, the focus will be limited primarily to an examination of the effect of income and education under different model specifications. Income and education are both substantively important in explaining differential mortality and both have been proposed as having nonlinear and nonadditive effects (Hadley, 1982; Adamchak and Robinson, 1986).

[Table 1 about here]

The coefficient on income suggests that, ceteris paribus, when the level of income in a community increases by $1,000, the mortality rate declines by .026 deaths per one thousand population. Similarly, a one year increase in median education is accompanied by a mortality decrease of .131 deaths per one thousand population. The impact of each variable can be appropriately evaluated as a unique and independent hypothesis by examining the standard t for the hypothesis that $\hat{\beta}_j = 0$.

In the present case, both coefficients are statistically significant beyond the .001 level. Since the five independent variables are obviously correlated among themselves, a persistent issue is to try and assess whether the degree of collinearity of the set is sufficiently pronounced to cause concern about the quality and reliability of the regression estimates. Following the strategy of Belsley et al. (1980), the issue can be addressed in the following hierarchical fashion[2]: (1) Determine the existence and intensity of the collinearity problem(s) (there can be more than one) among the independent variables, and (2) Identify the specific variables involved in each collinear situation and determine which of the estimated coefficients have been potentially degraded or adversely affected[3]. This can be accomplished with the help



BEST COPY AVAILABLE

of two pieces of diagnostic information (1) The condition numbers, $\kappa$, and (2) The regression coefficient variance decomposition values, $\phi$. Both pieces of information can be derived from the information generated when solving a system of linear equations with the eigen system[4].

Although the published literature in the area of collinearity diagnostics tends to be technical, for the current discussion it will suffice to point out that collinearity is indicated by the presence of a small eigenvalue of the cross product matrix of X'X when the independent variables have been scaled to have unit length (i.e. each observation is divided by the standard deviation). Specifically, Belsley et al. (1980) developed a condition index, $\kappa$, defined as the square root of the ratio of the largest eigenvalue of the scaled X'X matrix to each of the K-1 remaining eigenvalues[5] (i.e. $\kappa = (\lambda_1/\lambda k)^{1/2}$). The index measures how sensitive the regression parameters are to changes in the cross products matrix. The larger the condition number the more ill conditioned the data matrix and the more severe the collinearity problem. The guide lines presented by Belsley et al. (1980) suggest that condition indices of 30 or higher indicate levels of collinearity that potentially degraded parameter estimates. Clearly, there can be more than one problematic near dependency among the columns of the matrix of independent variables as indicated by the simultaneous presence of more than one large condition number. In fact, there can theoretically be K-1 near dependencies where K refers to the number of independent variables in the estimated equation, including the constant term.

Although the condition numbers identify the existence and number of potential collinearity problems, they do not provide information on the



BEST COPY AVAILABLE

nature and extent of the effect of such problems on the quality of the parameter estimates. This can be accomplished by decomposing the estimated variance of each regression coefficient, $\hat{\sigma}^2_{\beta_k}$, into additive components, each of which is associated with a single eigenvalue. Hence, if there are seven variables in the estimated equation, there will be seven eigenvalues, and each of the regression coefficient variances can be decomposed into seven terms[6].

Since the eigenvalues, $\lambda$, appear in the denominator of the variance decomposition equation (see Footnote 6), the variance proportions, $\phi_{kj}$, associated with small values of $\lambda_j$ will be large relative to other components making up the total variance of the coefficient. If one eigenvalue is associated with large variance proportions, for <u>two or more</u> independent variables (i.e. $\phi_{kj} \geq .50$) this indicates that collinearity is degrading the precision or increasing the variance of the regression coefficients for the designated variables. In this context, it should also be noted that not all coefficient variances are necessarily adversely affected by the presence of collinearity. If the coefficient for a specific independent variable does not have a significant amount (i.e. $\phi \geq .5$) of its variance associated with a small eigenvalue (i.e. a condition number$\geq$30), then that coefficient is unaffected by the existing collinearity. This regression variance decomposition approach is a significant improvement over the earlier strategies of examining a correlation matrix since it allows the researcher to pinpoint those parameters which are being adversely affected by collinearity. If theoretical interest centers only on the unaffected parameters, no problem exists and no further corrective action is necessary to produce reliable estimates (Judge, et al., 1985).


BEST COPY AVAILABLE

A further point is important in this context. Collinearity adversely effects OLS regression parameters by inflating the magnitude of regression coefficient variances, thus introducing imprecision into the results. If interest is limited to whether an adversely affected coefficient is significantly different from zero and the results of the t-test indicate that it is, there is no reason for concern or additional action (Belsley, et al., 1980). In short, collinearity results in a conservative bias (i.e. wide confidence intervals) in evaluating tests of statistical significance. With that in mind, we return to the collinearity diagnostics for the empirical model estimated in Equation 1.

Collinearity Diagnostics For The Community Mortality Model

We note again that the fundamental issue with collinear data is one of lack of independence of information. Variables that are collinear do not provide information that is meaningfully different from the information provided by other variables. When such near dependencies exist in the data it becomes problematic to accurately measure the separate influence of the affected variables on the dependent variables. To what extent do collinearity problems degrade the regression estimates for the linear Community Mortality Model? By examining the information produced in the collinearity diagnostics table (Table 2) it can be seen that there is one near dependency that is of concern. The variance of the coefficients for median education ($\phi$=.97) and the intercept term ($\phi$=.92) are associated with a condition index over 30. Hence, those two regression estimates have been degraded to some extent. However, since the confidence intervals are small and both are statistically

BEST COPY AVAILABLE

significant beyond the .001 level, it would appear that the degradation has not been harmful and confidence about the reliability of the parameter estimates is indicated. Further, it should be pointed out that, contrary to much of the thinking in sociology, income is not involved with education to produce a collinearity problem. In fact, only 8% of the coefficient variance for income is associated with the problematic condition index of 41.

[Table 2 about here]

The results of the linear model presented above were easily interpreted, and although there is one near dependency in the data affecting the intercept and the coefficient on education, the magnitude did not harmfully degrade the regression estimates. Turn now to an examination of a polynomial model that includes a quadratic term for income. The estimated model is:

$$\text{EQ2} = \text{mortality} = \hat{\beta}_0 + \hat{\beta}_1(\%\ \text{Nonwhite}) + \hat{\beta}_2(\text{Income}) + \hat{\beta}_3(\text{Education} + \hat{\beta}_4(\%\ \text{Urban}) + \hat{\beta}_5(\%\ \text{White collar}) + \hat{\beta}_6(\text{Income}^2) + \varepsilon$$

The OLS parameter estimates for Equation 2 are contained in column 2 of Table 1. As in the simple linear model, the interpretation of the effects of education is summarized in a straightforward manner by the coefficient $\hat{\beta}_3 = -.141$, implying an expected reduction in mortality of .141 deaths per 1,000 population for an increase in median education of one year. But what is the effect of income? By adding a quadratic term we have explicitly hypothesized that the effect of income is no longer constant for all levels of income. Instead, we have hypothesized that the effect of income on mortality depends on the level of income. Hence, the correct test of the effect of income is no longer


BEST COPY AVAILABLE

accomplished by testing $\hat{\beta}_2$ for significance. The appropriate test will necessarily consider simultaneously the coefficients of income, $\hat{\beta}_2$, and the square of income, $\hat{\beta}_6$. From the rules of calculus, the effect of the unit change in income on mortality is measured by the partial derivative:

$$\text{EQ3:} \quad \frac{\partial \text{ Mortality}}{\partial \text{ Income}} = \hat{\beta}_2 + 2\hat{\beta}_6(\overline{\text{income}})$$

The last term in the partial derivative shows explicitly that the expected change in mortality for a $1,000 change in income will depend on where along the income distribution the $1,000 change takes place. Table 3 presents calculated values for the effect parameter in Equation 3 for income values around the average median income for counties in the United Stated and values ±2 standard deviations from the mean. Notice first that the partial derivative does not turn negative until median income approaches $9,000, after which it decreases at an increasing rate. Hence, moving the median family income from $10,000 to $11,000 produces an expected mortality decline of -.005 deaths per thousand population. At $24,000 the same $1,000 increase produces an expected decline of -.042 per thousand.

As discussed above, the partial derivatives of quadratic models have a straightforward interpretation as a rate-of-change parameter. There are, however, two further questions that need to be addressed: (1) What is the effect of adding a quadratic term on the presence and levels of collinearity, and (2) What is the appropriate test statistic for the hypotheses on the effect of income? By looking at the second panel in Table 2, the impact of the inclusion of the squared income term

 BEST COPY AVAILABLE

on collinearity becomes clear. There are now two condition indexes over 30 involving four of the seven parameters of the model. The parameters of income ($\phi$=.99) and income$^2$ ($\phi$=.97) have virtually all of their variance affiliated with the condition index of 98. Further, the variance for the intercept term is also materially affected by the dominant near dependency ($\phi$=.53). Additionally, the second near dependency with the condition index of 44 accounts for 87 percent of the variance for the education parameter, as well as 46 percent of the variance for the intercept. Hence, a minimum of 97 percent of the variance of four coefficients is associated with two near dependencies suggesting potentially significant degradation of those regression estimates. The nature of the degradation is illuminated by noting what happened to the standard error for income when the income squared term was added to the model. In the linear model, the standard error for income was .0067, but in the quadratic model the standard error increased to .035, a 422 percent increase. However, the standard error on education was virtually unaffected, increasing only slightly from .021 to .022. Clearly the collinearity has introduced a good deal of imprecision into the parameter estimates. The impact on tests of significance are discussed below.

The data just discussed suggested that the inclusion of the quadratic term resulted in an increase in the levels of collinearity and inflation of the variances of the coefficients involving income. As a result, the regression coefficients on income and income squared are not statistically significant at any conventional level suggesting that income does not have an effect on mortality. However, as noted above, when a quadratic term is included in the model it is no longer correct

to test hypothesis on individual coefficients. Since the partial derivative measures the hypothesized effect of income, the correct test will involve testing the synthetic term directly.

The sampling variance necessary to test a partial derivative for statistical significance can be derived from the variance-covariance matrix of the regression coefficients and the results will vary depending on what level of income is being evaluated[6]. The probability levels for the partial derivatives are included in Table 3. It should be noted that even though the coefficients for income and income squared in Table 1 are not significant when tested independently, the derivative is significantly different from zero for values of income at or above the mean. Hence, the conclusions to be drawn will be very different if the derivative is tested directly.

The Multiplicative Model

The literature suggests that the effect of income on mortality will be conditioned on the level of education.' The formulation of such an expectation would be exactly as in Equation 2 except the income squared term is replaced by the multiplicative term: (Income)(Education). The partial derivative for income in the multiplicative model is:

$$EQ4: \quad \frac{\partial \text{Mortality}}{\partial \text{Income}} = \hat{\beta}_2 + \hat{\beta}_6 \text{Ed}$$

The nature of the conditioning effect is specified in the second term of Equation 4. An examination of Table 3 shows just how pronounced the interaction is. When education is at 9.5 years, a $1,000 increase in median family income results in an increase of .02 deaths per 1,000

 BEST COPY AVAILABLE

population. The sign on the effect parameter stays positive until the education level approaches 11 years, essentially the average for all counties in the United States. With education at 11.5 years, an increase in median family income of one thousand dollars is expected to lower the mortality rate by .018 deaths per 1,000 population and if education is at 13.5 years the effect is -.05. Clearly, to suggest that the impact of income is -.026 (the parameter in the linear model) is a significant distortion of what the conditioned relationship actually is. Note, however, that by including the multiplicative term the maximum condition index in Table 2C has increased to 192 and is associated with at least 90 percent of the variance of four of the seven parameters in the model, suggesting potential degradation of the very parameters of concern. Despite the imprecision introduced into the results, the tests of the derivatives are, nonetheless, statistically significant at .05 or beyond with the exception of the case where the educational level is 10.5 years.

The Full Model

Finally, it is possible to include both the nonadditive feature of the income education interaction and the nonlinear feature of income into one model as[8]:

$$\text{EQ5: Mortality} = \hat{\beta}_0 + \hat{\beta}_1(\%\ \text{Nonwhite}) + \hat{\beta}_2(\text{Income}) + \hat{\beta}_3(\text{Education}) + \hat{\beta}_4(\%\ \text{Urban}) + \hat{\beta}_5(\%\ \text{White Collar}) + \hat{\beta}_6(\text{Income}^2) + \hat{\beta}_7(\text{Income})(\text{Education}) + \varepsilon$$

The effect of income is now summarized by:

$$\text{EQ6: } \frac{\partial y}{\partial \text{Income}} = \hat{\beta}_2 + 2\hat{\beta}_6(\text{Inc}) + \hat{\beta}_7(\text{Ed})$$

The second two terms in Equation 6 make it clear that the effect of income will depend simultaneously on what level of income and what level of education is present in the community. Table 3 presents values for the partial derivative evaluated at the mean for education and income and one and two standard deviations around the mean values. Hence, if the income level is $10,000 and the educational level is 9.5 years, the effect of income is -.003 and is not statistically significant. At the mean levels of $17,000 and 11.5 years of education the effect is -.021, approaching the value in the linear model. When the income is at $24,000 and education is at 13.5 years the effect has increased to -.038. By evaluating the derivative for differing levels of income and education, it is possible to see the relative importance of the two variables. For example, if income was at $24,000 but education was at 9.5 years, the effect of income is an estimated increase in mortality of .067 deaths per thousand population ($P<.01$). Conversely, if the situation is reversed and the area has an income of only $10,000 but an average educational level of 13.5 years, the income effect is -.109 ($P<.01$). In short, the desirable effect of education more than outweighs the deleterious effect of income of low levels.

The last panel of data in Table 2 show that the magnitude of the largest condition index has increased to 226 and there are now a total of three near dependencies. However, the potentially degraded parameters remain the same although the variance proportions are distributed somewhat differently. Specifically, the variance of income and the square of income is distributed more evenly across the two highest condition indices while the variance for education and the

interaction of income and education is still dominated by the largest index.

Summary and Implications

The paper examined how the inclusion of quadratic and/or multiplicative terms affected the levels of collinearity and the precision of the parameter estimates in a model estimated with ordinary least squares, and how the effects of the germane variables can be measured and tested for statistical significance. Adding a quadratic term to the linear model increased the largest condition index from 41 to 98. By including a multiplicative term the index increased to 192 and for the full model the index reached 226. Belsley et al. (1980) point out that condition numbers increase roughly along the scale of magnitude 10, 30, 100, 300, 1000. Hence, the inclusion of each separate synthetic term increased the magnitude of the near dependencies by roughly one level over the linear model, although the condition number for the multiplicative model was somewhat larger. Further, the simultaneous inclusion of the two terms did not significantly increase the level of ill conditioning beyond what each produced individually.

The results of the collinearity diagnostics confirm that as many as four of the regression coefficients in the multiplicative and full models had estimated variances determined principally by the presence of a highly collinear relation and were thus degraded. The degradation was the result of the collinearity induced by including nonadditive terms into the equation. However, tests of the partial derivatives indicated that despite the increased coefficient variances produced by the relatively high levels of collinearity, most were statistically


BEST COPY AVAILABLE

significant at the .05 level or beyond. Hence, in the present case the problematic issue revolves around the use of relatively imprecise point estimates and wide confidence intervals. If the coefficient variances become too large, interval estimates of the coefficients for policy formulation or forecasting become useless, even if the coefficients are statistically significant. In such instances it would be necessary to somehow deal with the harmful levels of collinearity. The easiest way to solve the problem is to drop highly collinear variables from the model. However, if the model specification is theoretically correct, dropping variables will result in misspecification. In many instances the deleterious effects of misspecification can be more problematic than those of collinearity. An alternative strategy for dealing with the collinearity problem is to re-estimate the model with an estimator which is designed to cull more information from the sample than is possible with ordinary least squares and thus producing more precise parameter estimates. The three that are most widely accepted are the principal components regression, the stein estimator, and the ridge regression estimator. As Judge et al. (1985) point out, there are drawbacks and limitations to these alternatives, and in general, they are only useful in situations where interval estimates are unnecessary and point estimates are desired since the sampling distributions of the estimators are not yet understood.

Finally, several comments about the levels of collinearity in the various model is in order. The prevailing sentiment tends to be that collinearity in structural social science models is rampant, particularly in models which include variables, such as education, income, occupational structure, and race. There was, in fact, some

modest degree of collinearity in the linear model which included all of these variables, but collinearity essentially caused no problem in the interpretation of results. When quadratic and/or multiplicative interaction terms were included in the model the level of collinearity did increase. However, even in those instances the effects of collinearity were localized and entailed only those variables with a common base. Hence, there is some trade off between estimating models that contain theoretically appropriate nonadditive and nonlinear terms and resultant methodological problems. However, the substantive insight gained from including the nonlinear and nonadditive terms in regression models appears to outweigh the increased complexity of measuring and statistical testing of causal effects and the increased problems of collinearity.

BEST COPY AVAILABLE

Table 1
Parameter Estimates for Regression Models

(Probability Levels for t in Parenthesis)

| Variable | Model Form | | | |
|---|---|---|---|---|
| | Linear | Quadratic | Multiplicative | Quadratic and Multiplicative |
| Constant | 9.39 (<.001) | 9.10 (<.001) | 6.44 (<.001) | 6.11 (<.001) |
| % Nonwhite | .030 (<.001) | .030 (<.001) | .029 (<.001) | .029 (<.001) |
| Income | -.026 (<.001) | .022 (.529) | .214 (<.001) | .197 (<.001) |
| Education | -.131 (<.001) | -.141 (<.001) | .115 (.03) | .211 (<.002) |
| % Urban | .001 (.11) | .001 (.119) | .001 (.06) | .001 (.04) |
| % White Collar | -3.37E-4 (.59) | -2.36E-4 (.709) | 5.54E-7 (.99) | -8.39E-5 (.89) |
| $Income^2$ | — | -.001 (.160) | — | .002 (.04) |
| Income x Education | — | — | -.020 (<.001) | -.026 (<.001) |
| $\bar{R}^2$ | .233 | .233 | .239 | .240 |

Table 2
Collinearity Diagnostics: Eigenvalues, Condition Indices, and Variance-Decomposition Proportions

Linear Model

Variance Decomposition Proportions

| Eigenvalue | Condition Index | Intercept | % Black | Medinc | Med | % Urban | % White Collar |
|---|---|---|---|---|---|---|---|
| 4.73 | 1.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 |
| 0.72 | 2.56 | 0.00 | 0.81 | 0.00 | 0.00 | 0.01 | 0.01 |
| 0.29 | 4.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.00 |
| 0.23 | 4.49 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.98 |
| 0.02 | 16.58 | 0.07 | 0.08 | 0.91 | 0.02 | 0.23 | 0.00 |
| 0.00 | 41.12 | 0.92 | 0.09 | 0.08 | 0.97 | 0.01 | 0.00 |

Quadratic Model

Variance Decomposition Proportions

| Eigenvalue | Condition Index | Intercept | % Black | Medinc | Med | % Urban | % White Collar | Incsq |
|---|---|---|---|---|---|---|---|---|
| 5.64 | 1.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 |
| 0.75 | 2.74 | 0.00 | 0.79 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.29 | 4.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.00 | 0.00 |
| 0.25 | 4.78 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.97 | 0.00 |
| 0.07 | 8.74 | 0.01 | 0.11 | 0.00 | 0.01 | 0.26 | 0.01 | 0.03 |
| 3.0E-3 | 44.86 | 0.46 | 0.09 | 0.01 | 0.87 | 0.00 | 0.00 | 0.00 |
| 6.0E-4 | 98.39 | 0.53 | 0.01 | 0.99 | 0.12 | 0.01 | 0.01 | 0.97 |

Multiplicative Model

Variance Decomposition Proportions

| Eigenvalue | Condition Index | Intercept | % Black | Medinc | Med | % Urban | % White Collar | (Inc)(Ed) |
|---|---|---|---|---|---|---|---|---|
| 5.69 | 1.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 |
| 0.74 | 2.77 | 0.00 | 0.79 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.30 | 4.39 | 0.00 | 0.00 | 0.00 | 0.00 | 0.69 | 0.01 | 0.00 |
| 0.24 | 4.83 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.97 | 0.00 |
| 0.03 | 13.63 | 0.01 | 0.13 | 0.00 | 0.00 | 0.23 | 0.00 | 0.01 |
| 4.0E-3 | 38.93 | 0.05 | 0.07 | 0.03 | 0.09 | 0.00 | 0.00 | 0.01 |
| 1.0E-4 | 192.46 | 0.94 | 0.00 | 0.97 | 0.91 | 0.00 | 0.01 | 0.98 |

Full Model: Multiplicative and Quadratic

Variance Decomposition Proportions

| Eigenvalue | Condition Index | Intercept | % Black | Medinc | Med | % Urban | % White Collar | (Inc)(Ed) | Incsq |
|---|---|---|---|---|---|---|---|---|---|
| 6.61 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 |
| 0.76 | 2.94 | 0.00 | 0.76 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.30 | 4.73 | 0.00 | 0.00 | 0.00 | 0.00 | 0.69 | 0.01 | 0.00 | 0.00 |
| 0.25 | 5.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.96 | 0.00 | 0.00 |
| 0.08 | 9.26 | 0.00 | 0.12 | 0.00 | 0.00 | 0.25 | 0.01 | 0.00 | 0.01 |
| 4.0E-3 | 38.02 | 0.06 | 0.10 | 0.01 | 0.02 | 0.00 | 0.00 | 0.02 | 0.03 |
| 6.0E-4 | 100.85 | 0.03 | 0.00 | 0.38 | 0.09 | 0.01 | 0.01 | 0.03 | 0.76 |
| 1.0E-4 | 226.92 | 0.91 | 0.01 | 0.62 | 0.89 | 0.00 | 0.00 | 0.95 | 0.20 |

Table 3
Partial Derivatives of Mortality with Respect to Income
(Probability levels for t are in parentheses)

---

Linear Model

$$\frac{\partial \text{Mortality}}{\partial \text{Income}} = \hat{\beta}_2 = -.026$$

Quadratic Model

$$\frac{\partial \text{Mortality}}{\partial \text{Income}} = \hat{\beta}_2 + 2\hat{\beta}_6 \text{ (Income)}$$

| Income Level | | Parameter Value |
|---|---|---|
| $-2\sigma$ | \$10,000 | -.005 (.79) |
| $-1\sigma$ | \$13,000 | -.013 (.29) |
| $\bar{X}$ | \$17,000 | -.023 (<.01) |
| $+1\sigma$ | \$20,000 | -.032 (<.01) |
| $+2\sigma$ | \$24,000 | -.042 (<.01) |

Multiplicative Model

$$\frac{\partial \text{Mortality}}{\partial \text{Income}} = \hat{\beta}_2 + \hat{\beta}_6 \text{ (Ed)}$$

| Education Level | | Parameter Value |
|---|---|---|
| $-2\sigma$ | 9.5 | .022 (.05) |
| $-1\sigma$ | 10.5 | .002 (.77) |
| $\bar{X}$ | 11.5 | -.018 (.01) |
| $+1\sigma$ | 12.5 | -.038 (<.01) |
| $+2\sigma$ | 13.5 | -.058 (<.01) |

Multiplicative and Quadratic Model

$$\frac{\partial \text{Mortality}}{\partial \text{Income}} = \hat{\beta}_2 + 2\hat{\beta}_6 \text{ (Inc)} + \hat{\beta}_7 \text{ (Ed)}$$

| Income Level | | Education Level | Parameter Value |
|---|---|---|---|
| $-2\sigma$ | \$10,000 | 9.5 | -.003 (.85) |
| $-1\sigma$ | \$13,000 | 10.5 | -.014 (.22) |
| $\bar{X}$ | \$17,000 | 11.5 | -.021 (.01) |
| $+1\sigma$ | \$20,000 | 12.5 | -.032 (<.01) |
| $+2\sigma$ | \$24,000 | 13.5 | -.038 (<.01) |

FOOTNOTES

2. The collinearity diagnostics is approached from the perspective proposed by Belsley et al. (1980) since that procedure or its equivalent is now readily available as an option in many of the commercial statistical packages.

3. It should be explicitly noted that collinearity problems are <u>data</u> problems, not statistical problems. The issue has to do with specific characteristics of the data matrix, X, not with the OLS regression model. However, the data problems are of considerable moment to the efficacy of OLS estimation.

4. The strategy involves deriving estimates for the regression coefficients by finding the characteristic roots an vectors (i.e. eigenvalues and eigenvectors) of the moment matrix, X'X. The strategy involves setting up a system of homogeneous equations of the form:

$$(A - \lambda I)K = 0$$

with the elements of K as the unknowns to be solved for in addition to the unknown value of $\lambda$, the eigenvalues for the matrix A. The next step is to solve for the values of $\lambda$ that will produce a determinant for the system that is zero. Specifically:

$$|A - \lambda I| = 0$$

This turns out to be an $K^{th}$ degree polynomial in $\lambda$, which in general will have K different roots.

It is useful to note that a conventional structural equation (i.e. a non-homogeneous equation) can be transformed into a homogeneous equation by the use of a dummy variable. Consider the general non-homogeneous equation formulation:

$$BX = K$$

Now consider a specific two equation system:

$$\begin{aligned} 2x_1 + 3x_2 &= 9 \\ 5x_1 - x_2 &= 14 \end{aligned}$$

$$\text{or} \quad \begin{pmatrix} 2 & 3 \\ 5 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 9 \\ 14 \end{pmatrix}$$

Set the two equations to zero and add a dummy, $x_3=-1$ to each to form the homogeneous equation of the form:

$$(B|K)X = 0$$

Specifically you will have:

$$\begin{aligned} 2x_1 + 3x_2 + 9x_3 &= 0 \\ 5x_1 - x_2 + 14x_3 &= 0 \end{aligned} \quad \text{or} \quad \begin{pmatrix} 2 & 3 & 9 \\ 5 & -1 & 14 \end{pmatrix} \begin{pmatrix} x1 \\ x2 \\ x3 \end{pmatrix} = 0$$

Hence, traditional structural equations can be thought of as a special case of homogeneous equations. The transformation is accomplished by adding a vector of constants, k, to the matrix B as an $(n+1)^{th}$ column (i.e. we added the column vector [9,14]', and augmenting the vector x with an element $x_{n+1}$, in this case $x_3$ which was set equal to -1.

5. Belsley et al. (1980) present their discussion in terms of singular value decomposition of a data matrix, X, and not the eigenvalues of a cross product matrix, X'X, noting that the singular values, $\mu$, are preferable since they can be calculated without recourse to an inverse. However, the notion of eigenvalues is more familiar to social scientists and, additionally, SAS collinearity diagnostics are reported with eigenvalues of X'X rather than singular values of X. The formal presentation of the issue would point out that the eigenvalues are squares of the singular values (i.e. $\lambda_k=\mu_k^2$).

6. Specifically, the variance of any given regression coefficient can be expressed as:

$$\hat{\sigma}^2_{\beta k} = \hat{\sigma}^2 \sum_{j=1}^{k} \frac{P^2_{kj}}{\lambda j}$$

The proportion of the variance of $\hat{\beta}_k$ associated with any particular eigenvalue will then be given by:

$$\phi_{kj} = \frac{P^2_{kj}/\lambda_j}{\sum_{j=1}^{K} P^2_{kj}/\lambda_j}$$

7. The sampling variance of a partial derivative is:

$$\sigma^2_{\partial y/\partial x} = \sum_{i=1}^{n} K_i(\sigma_i^2) + \sum_{i=1}^{n}\sum_{j\neq i} k_i k_j \operatorname{cov}(\hat{\beta}_i, \hat{\beta}_j)$$

Where $\sigma_i^2$ = sampling variance of $\hat{\beta}_i$

$K_i(1,2...n)$ = constants or weights on the parameters

$\operatorname{cov}(\hat{\beta}_i, \hat{\beta}_j)$ = sampling covariances of $\hat{\beta}_i$ and $\hat{\beta}_j$

Practically the standard errors for a partial derivative can be easily estimated by specifying a vector with the appropriate weighting constant (K) and then premultiplying the covariance matrix of the parameters by K' and postmultiplying the resultant product matrix by K. Hence, the appropriate constant vector for determining the sampling variance of income in Equation 2 is:

K' = [0,1,0,0,0,0,2 Inc] i.e. the weighting on $\hat{\beta}_2$ is 1 and the weighting on $\hat{\beta}_6$ is 2(Income)

And:

$$\sigma^2_{\partial y/\partial x} = \underset{\sim}{K}' \underset{\sim}{B} \underset{\sim}{K}$$

Where B is the covariance matrix of parameters that is standard output from most commercially available software packages.

Alternatively, the derivative can be tested for statistical significance by invoking the test procedure in SAS. For example, to test whether the partial derivative for income in the quadratic model in Equation 2 is significantly different from zero you specify TEST: $x_2 + 2 * x_6 * (\overline{Inc}) = 0$, where $(\overline{Inc})$ is the chosen level of income under consideration.

The information produced from the test is an F value and a probability level of F. The standard error can be imputed from the provided information as:

$$\hat{\sigma}_{\frac{\partial y}{\partial x}} = \frac{\hat{\beta}^*}{\sqrt{F}}$$

Where $\hat{\beta}^* = \frac{\partial y}{\partial x}$

8. An alternative specification could be:

$$\text{Mortality} = \hat{\beta}_1(\text{Education}) + \hat{\beta}_2(\text{Income}) + \hat{\beta}_3(\text{Income})(\text{Education}) + \hat{\beta}_4(\text{Education})(\text{Income})^2$$

In this formulation the extent to which income influences the effect of education varies nonlinearly over the range of income. The effect parameters for education and income would be:

$$\frac{\partial \text{Mortality}}{\partial \text{Education}} = \hat{\beta}_1 + \hat{\beta}_3(\text{Income}) + \hat{\beta}_4(\text{Income})^2$$

$$\frac{\partial \text{Mortality}}{\partial \text{Income}} = \hat{\beta}_2 + \hat{\beta}_3(\text{Education}) + 2\hat{\beta}_4(\text{Education})(\text{Income})$$

BEST COPY AVAILABLE

References

Adamchak, Donald J. and James L. Robinson
1986 "Evidence of A Curvalinear Relationship Between Income and Infant Mortality: Individual Data Do Not Bear Out Ecological Correlations." Sociology and Social Research 70(3):214-216.

Ahern, Mary C.
1979 Health Care in Rural America. Washington, D.C.: U.S. Department of Agriculture, Agriculture Information Bulletin No. 428.

Belsley, David A.
1982 "Assessing the Presence of Harmful Collinearity and Other Forms of Weak Data Through A Test for Signal-To-Noise." Journal of Econometrics 20:211-253.

Belsley, David A., Edwin Kuh, and Roy E. Welsch
1980 Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons.

Blalock, Hubert M.
1965 "Theory Building and the Concept of Interaction." American Sociological Review 30:374-381.

Dutton, D.B.
1984 "Socioeconomic Status and Children's Health." Medical Care 23(2):142-156.

Fuch, Victor R.
1974 Who Shall Live? New York: Basic Books, Inc.

Grossman, Michael
1973 "The Correlation Between Health and Schooling." Washington, D.C.: Conference on Research in Income and Wealth, Household Production and Consumption, National Bureau of Economic Research Report No.22.

Hadley, Jack
1982 More Medical Care Better Health? An Economic Analysis of Mortality Rates. Washington, D.C.: The Urban Institute Press.

Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee
1985 The Theory and Practice of Econometrics. New York: John Wiley and Sons.

Kitagawa, Evelyn M. and Philip M. Houser
1968 "Educational Differentials in Mortality by Cause of Death: United States 1960." Demography 5(1):318-353.

Mechanic, David
1968 Medical Sociology: A Selective View. New York: The Free Press.

BEST COPY AVAILABLE

SAS Institute, Inc.
1985 SAS User's Guide: Statistics. Version 5 Edition. Cary, North Carolina: SAS Institute, Inc.

Southwood, Kenneth E.
1978 "Substantive Theory and Statistical Interaction: Five Models." American Journal of Sociology 83(5):1154-1203.

Stokes, C. Shannon and Michael K. Miller
1985 "A Methodological Review of Fifty Years of Research in Rural Sociology." Rural Sociology 50(4):539-560.

Walker, Henry A. and Cohen Bernard P.
1985 "Scope Statements: Imperatives for Evaluating Theory." American Sociological Review 50:288-301.

BEST COPY AVAILABLE